

Assignments and grading

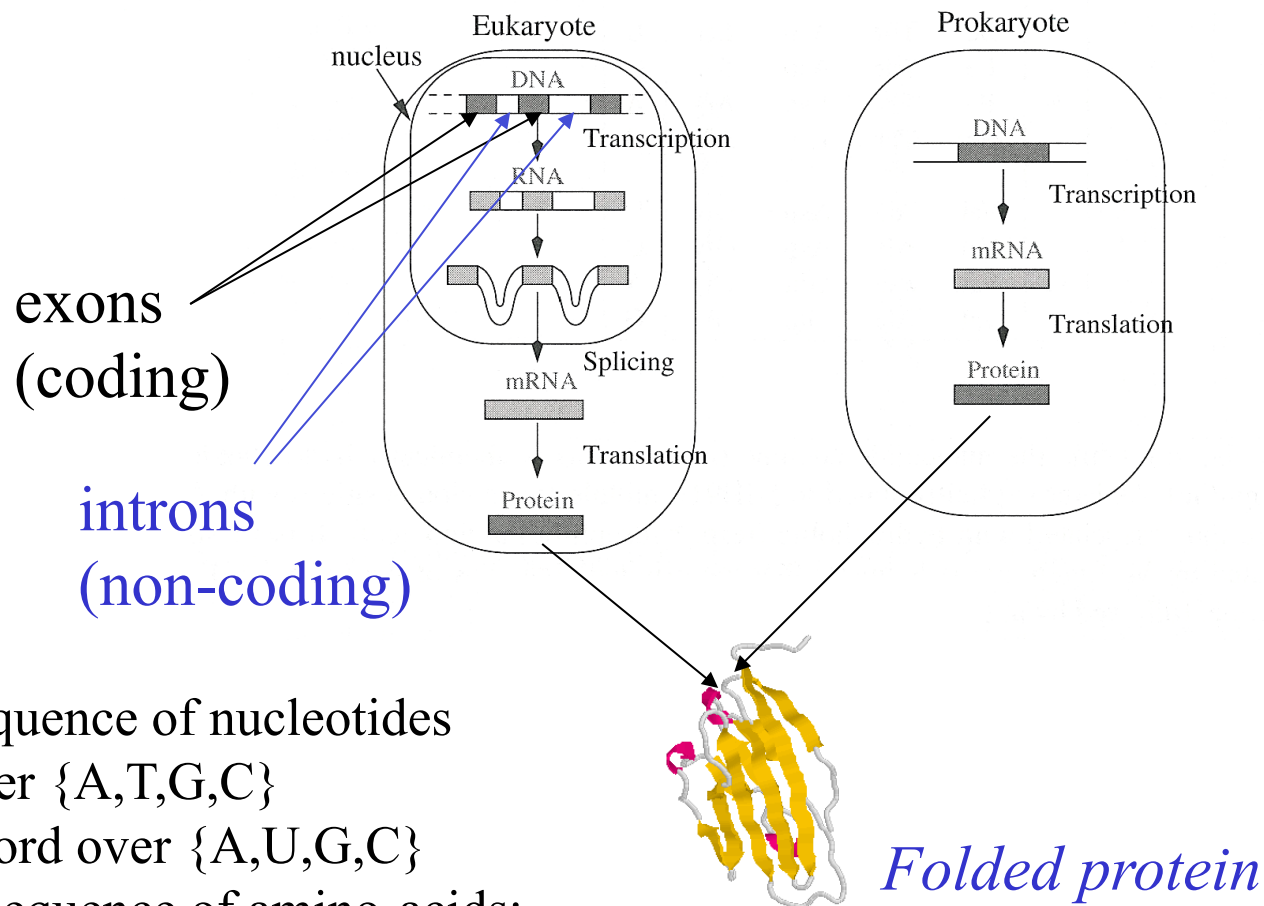
- Assignments and grading
 - Assignments (65%)
 - Class participation (5%)
 - Final exam (30%)
 - Grading – standard percentile
- Course info:
 - Lectures, homework etc.: course web page:
 - <http://www.apl.jhu.edu/~przytyck/>
 - Teresa.Przytycka@verizon.net
 - Igor
 - Best way for communication – e-mail;

Main Objective

- Understand computational techniques used in Computational Biology
- The focus is on methods not on learning particular software
-

Very Basic Facts about
organization of living
organisms

Organization of modern organisms



DNA-sequence of nucleotides

Word over $\{A, T, G, C\}$

RNA- word over $\{A, U, G, C\}$

Protein-sequence of amino-acids;
word over twenty letter alphabet
 $\{A, V, L, I, G, P, \dots\}$

DNA basics

Double-stranded: Single strand of DNA in one direction is paired to a complementary strand

5' AACTGC 3'

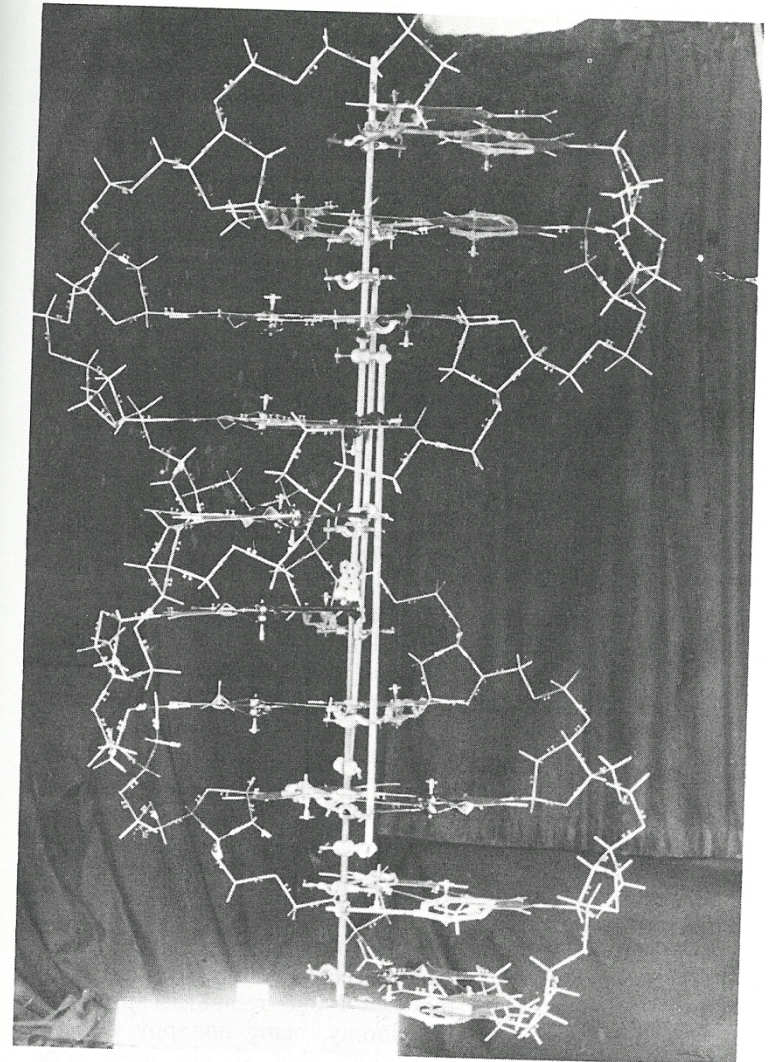
3' TTGACG 5'

forming in 3-dimension the double helix

DNA has associated with it directionality corresponding to the direction of translation.

DNA store information that encodes proteins

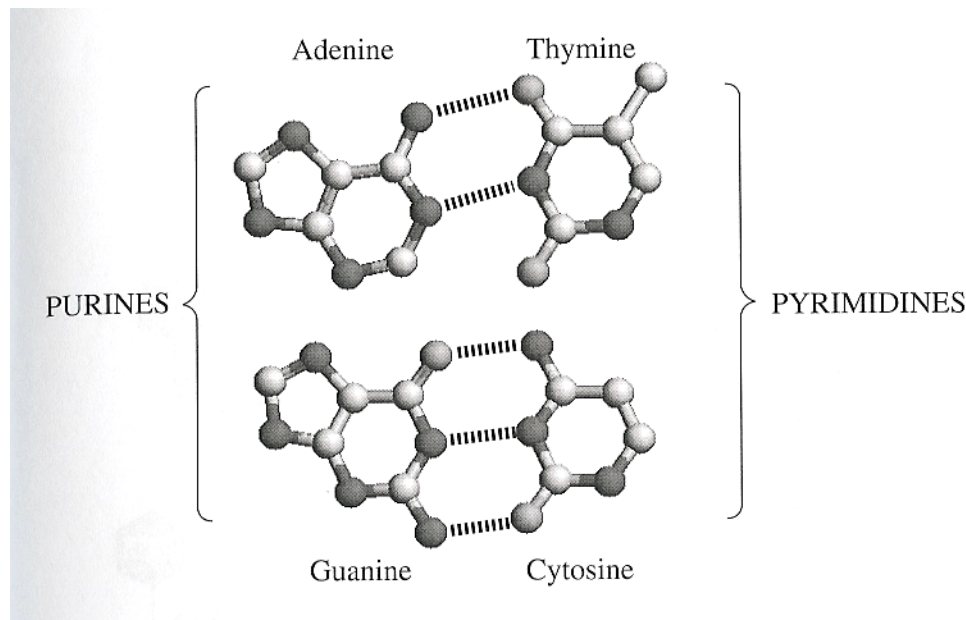
The Double Helix • 121



The original demonstration model of the double helix (the scale gives distances in Angstroms).

Base Pairs

- adenine (A), cytosine (C), guanine (G) and thymine (T).



Complementary **base pairs (bp)** (the so called Watson-Creek base pairs) : A-T, G-C

(Note the difference in the number of hydrogen bounds formed in each pair)

Hydrogen bond



- H-atom has a large positive charge. If its contact partner has a large negative charge hydrogen bond can form.
- The positive H^+ charges assumes it's lowest potential energy between two negative charges **when all three charges are linear.**
- The distance between hydrogen bonded atoms is smaller than it would follow from van der Waals radii.

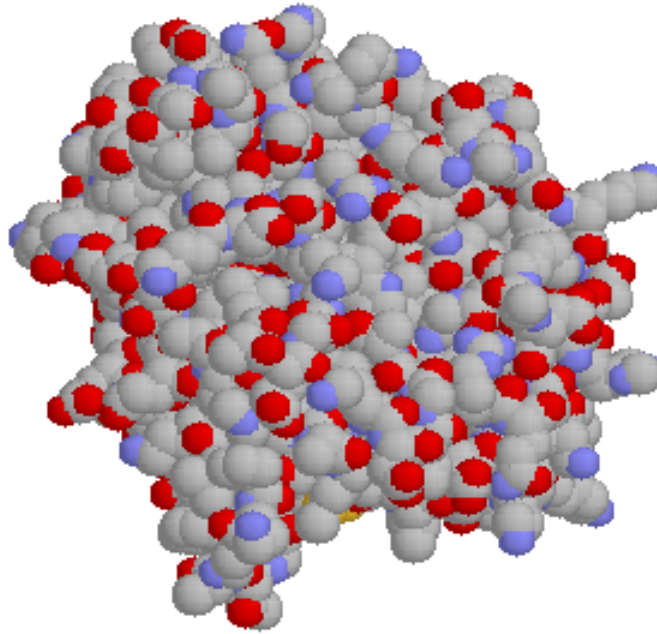
Example $\text{H} + \text{O} = 2.7\text{\AA}$

observed distance 1.9\AA

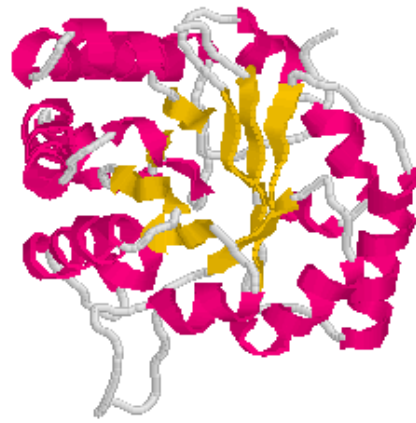
Proteins

- Polymers with 20 amino- acids as building blocks
- No complementary pairing
- Perform virtually all work in the organism: enzymes, transport, signaling

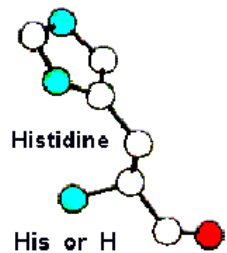
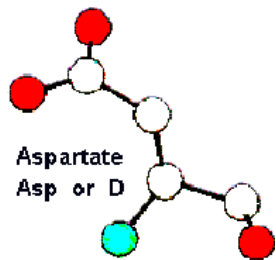
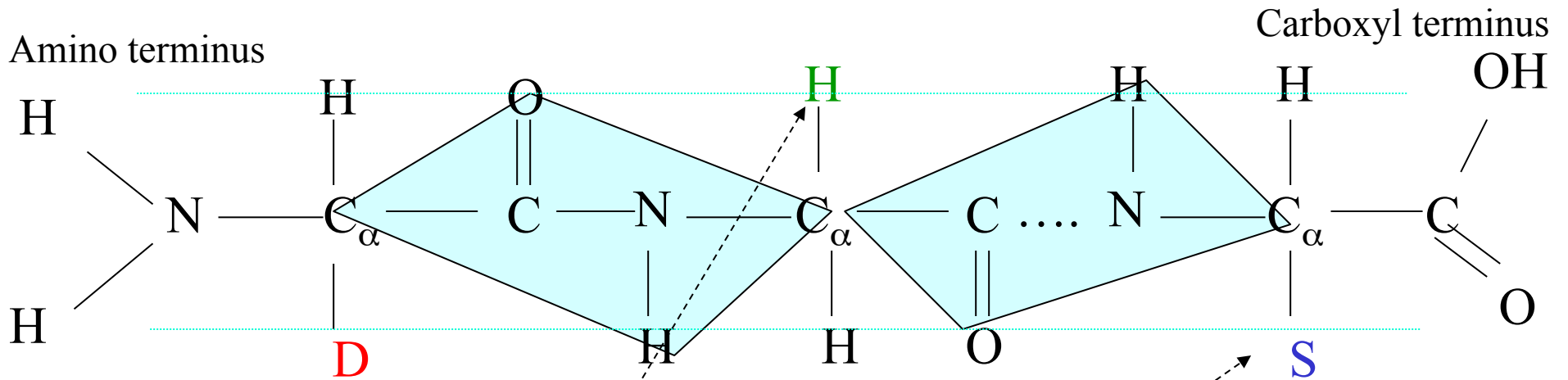
A protein in the space filling model



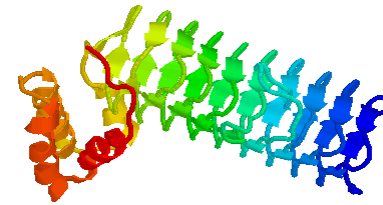
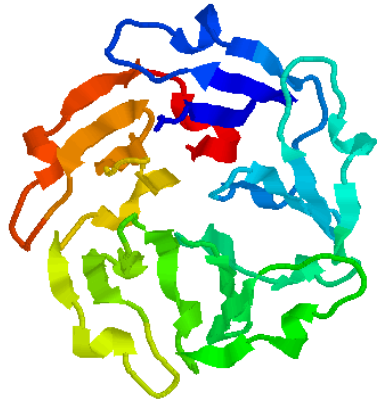
Ribbon representation



A closer look at a protein



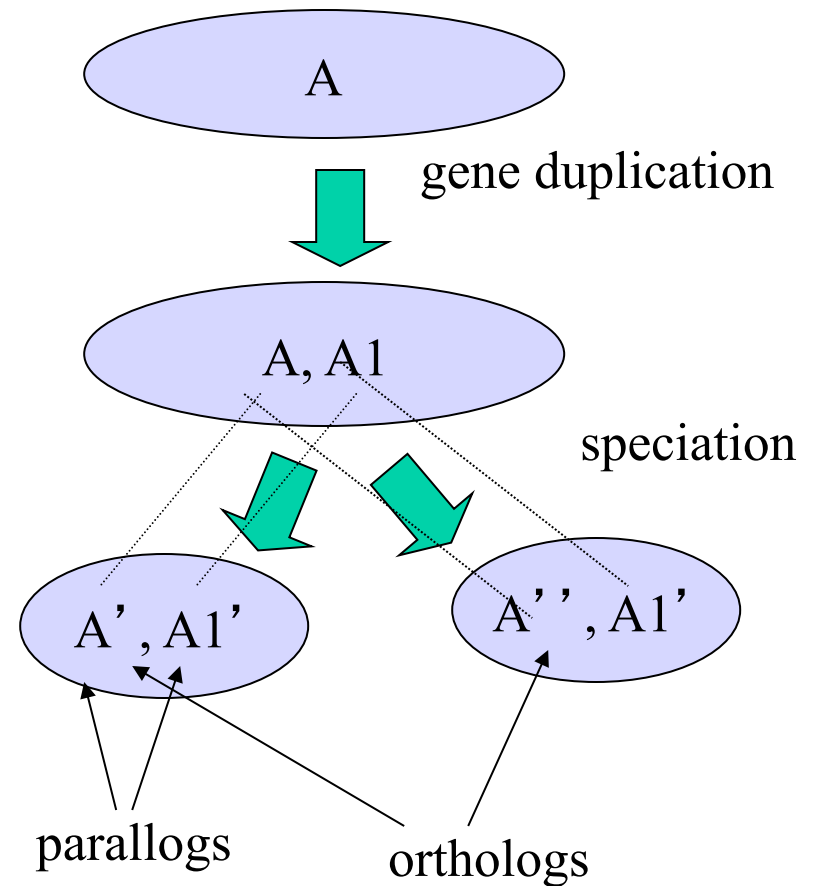
Proteins come in different shapes sizes, and numbers:



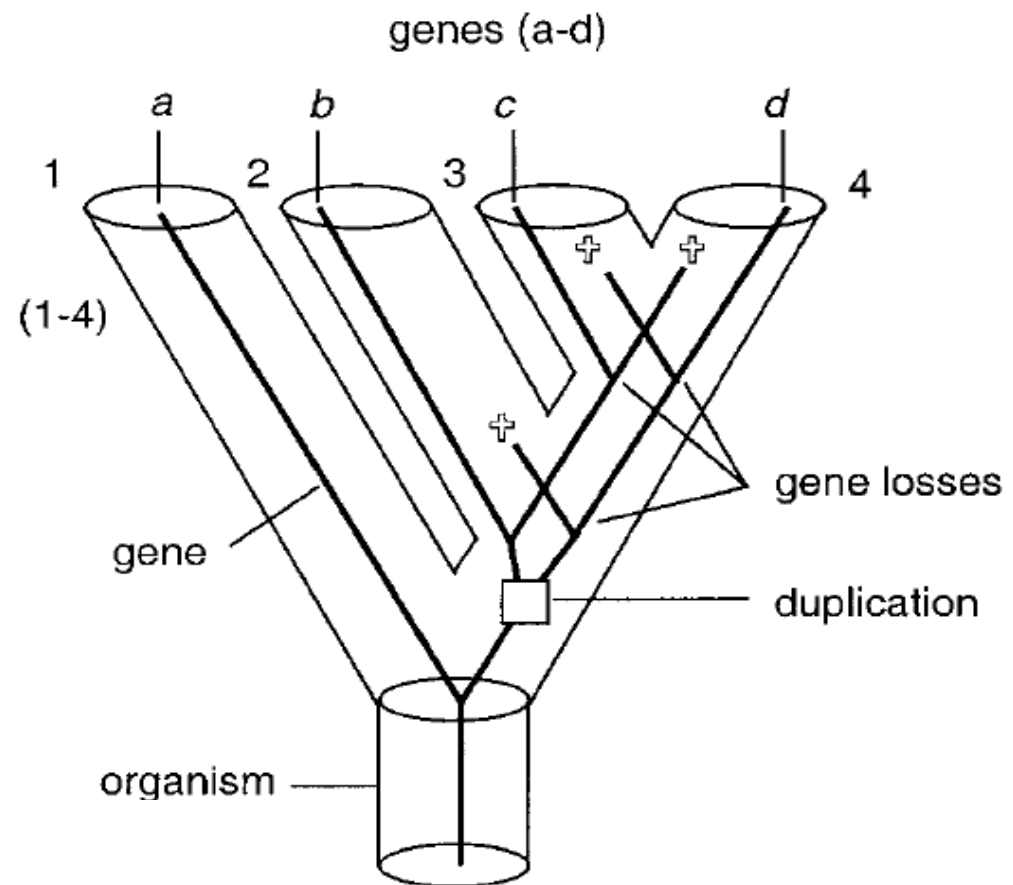
Evolutionary relations between genes

Types of evolutionary relationship

- **Homologs**: Sequences that descended from the same common ancestral sequence.
- **Orthologs** the least common ancestor was a speciation event (thus they must be in different species)
- **Paralogs**: A pair of genes from the same genome, the last common ancestor was a duplication event.



Gene tree and species tree



Molecular Phylogenetics and Evolution
Vol. 14, No. 1, January, pp. 89–106, 2000
Article ID mpev.1999.0676, available online at <http://www.idealibrary.com> on IDEAL®

Extracting Species Trees From Complex Gene Trees: Reconciled Trees And Vertebrate Phylogeny

Roderic D. M. Page

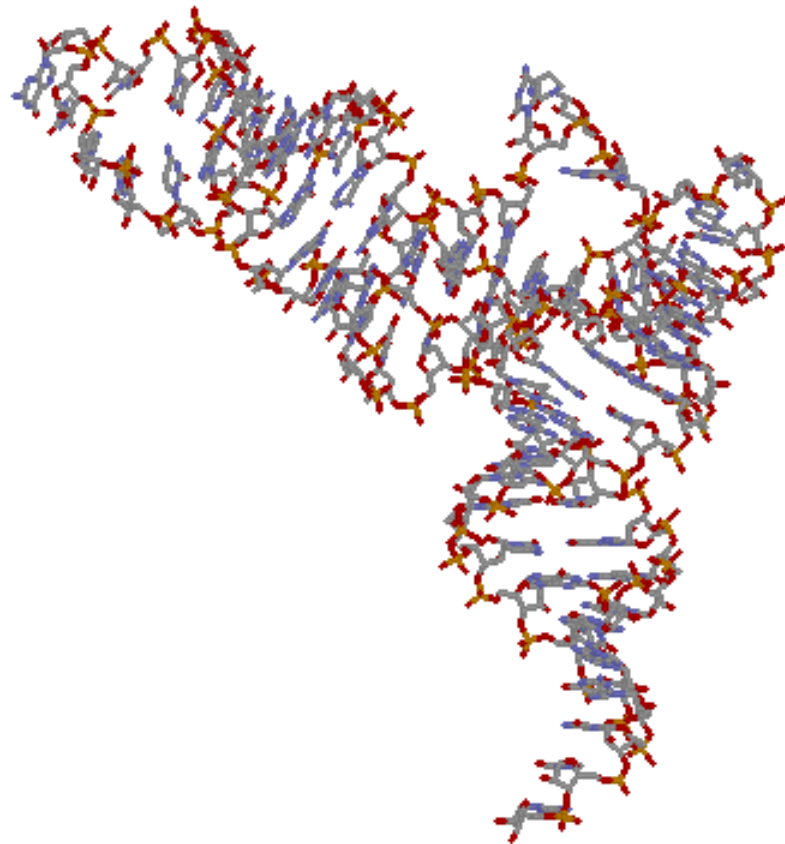
*Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences,
University of Glasgow, Glasgow G12 8QQ, United Kingdom*

Received December 14, 1998; revised April 22, 1999

RNA

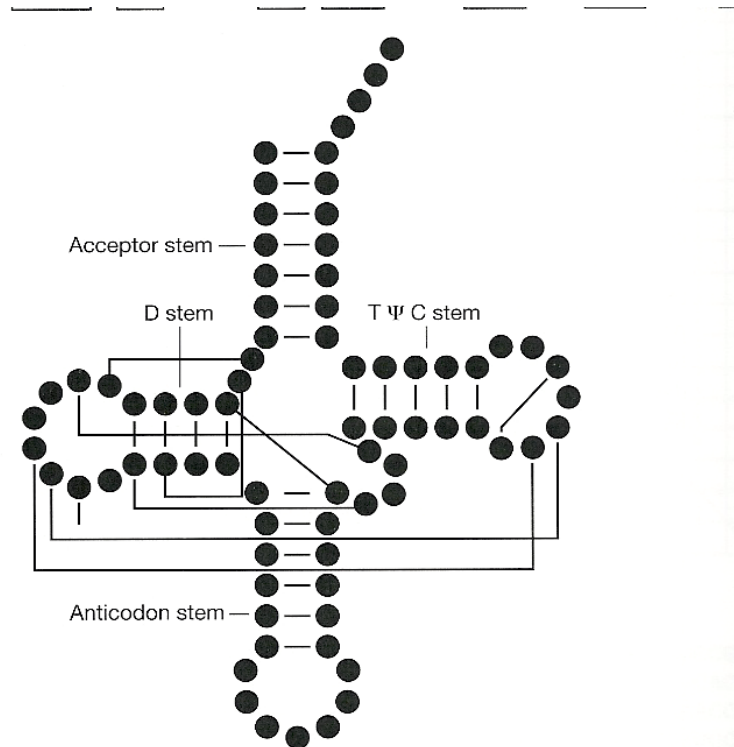
- Single stranded
- Thymine (T) is replaced by uracyl (U).
- Base-pairs are formed within the single strand.
The structure of base pairs = secondary tertiary structure; 3D arrangement
- Used as “temporary storage” for genetic information during translation (next slide)
- Can have catalytic activity (example of a job typical for a proteins) which prompts hypothesis about ancient “RNA world”

RNA structure



RNA secondary structure

B



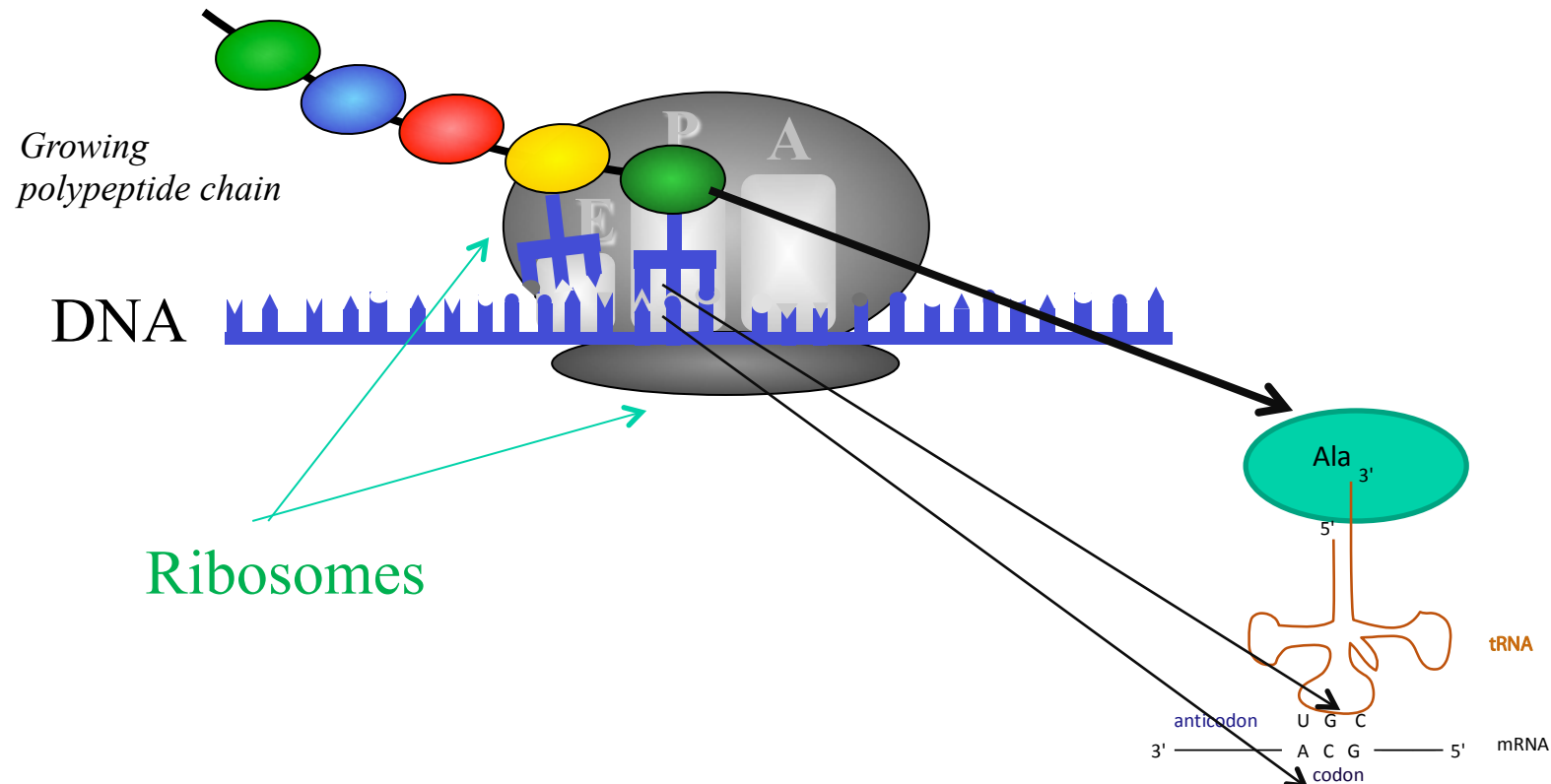
Genetic Code

- Amino-acids are encoded by triplets of nucleotides called codons
- Genetic code is “non-overlapping and comma free”
- The genetic code is redundant: there are 64 possible codons and 20 amino-acids + special “**stop**” codon.
- AUG (coding for Methionine) is “**start**” codon.

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Translation

Fundamental in translation is the so called transfer RNA (tRNA) molecules – linked to a specific amino-acid on one side and containing the triple complementary to the codon (the anticodon) on the other side.



Reading frames

A **reading frame** is one of three possible ways of grouping triples of nucleotides into codons.

- Example: ...TAAATAGAT...

Reading frames:

...,TAA,ATA,GAT,...

...T,AAA,**TAG**,AT. (TAG – stop codon)

...TA,AAT,AGA,T...

- An **open reading frame** (ORF) contiguous stretch of DNA beginning at the start codon, having an integral number of codons none of them being a stop codon. Open reading frame may contain a gene.

Example of a bioinformatic problem

- Find all genes in a genome.
- Why this is hard?
 - Genome contains coding and non-coding regions
 - You need to determine precisely where the gene starts (shifting by one nucleotide changes completely the encoded protein)
 - Find and determine precisely boundary of exons
 - How to tell “coding region” from “non-coding region”

Non-coding DNA

- Less than 5% on human code proteins (but more than 80% microbial genome is coding).
- More than 50% are various repeats:
 - Transposon-derived repeats (45%)
 - Inactive copies of cellular genes (pseudogenes)
 - Simple sequence repeats consisting of direct repetition of k-mers (eg, $(CA)_n$) (tandem repeats)
 - Segmental duplication consisting of blocks of DNA that have been copied from one place of genome to the other.

Part b

Basics of Information Theory

Probability distribution

Discrete probability distribution function, $p(x)$,
is a function on a discrete set of event that
satisfies the following properties.

- The probability that x can take a specific value is $p(x)$ ($p(x)$ is non-negative for all real x)
- The sum of $p(x)$ over all possible values of x is 1.

1. How probability distributions can be compared?
2. What the difference can tell us
3. Why biologists should care

Entropy: intuition

“measure of disorder”

- The maximum entropy – the item you are searching for has equal probability of being in any place
- Minimum entropy – with probability 1 you know where your item is

Entropy - formal definition

Entropy - function on probability distribution

$$P = p_1, p_2, \dots, p_n$$

Defined as the sum:

$$H(P) = -\sum_{i=1..n} p_i \log p_i$$

Max. entropy distribution is uniform distribution:

$$-\sum_{i=1..n} (1/n) \log (1/n) = \log n$$

Minimum entropy distribution $p = 1, 0, 0, \dots, 0$

$$H(p) = -\log 1 = 0$$

Information content

R- reference distribution (usually uniform distribution)

P – alternative distribution.

Information context $I(P)$ is defined as


$$I(P) = H(R) - H(P)$$

Intuition: The more distant a given probability distribution is from the uniform distribution (“most random distribution”) the more information it contains.

Example

Consider a set of homologous genes. Because of evolutionary changes, the sequences are not quite the same. Using multiple sequence alignment (later lectures) we can decide which nucleotide in one gene corresponds to which nucleotide in other genes:

Org 1	A	T	C	T	G	A	...
Org 2	A	T	A	C	G	A	...
Org 3	A	T	A	A	G	T	...
Org 4	A	T	A	G	G	T	...



Reference distribution: uniform (entropy 2)

Red column entropy = 0

Green column entropy = $-4(1/4)\log(1/4) = 2$

Blue column entropy = $-2(1/2)\log(1/2) = 1$

Red info content $2-0 = 2$

Green info content $2-2=0$

Blue info content $2-1 = 1$

Log base 2 = info content computed in **bits**

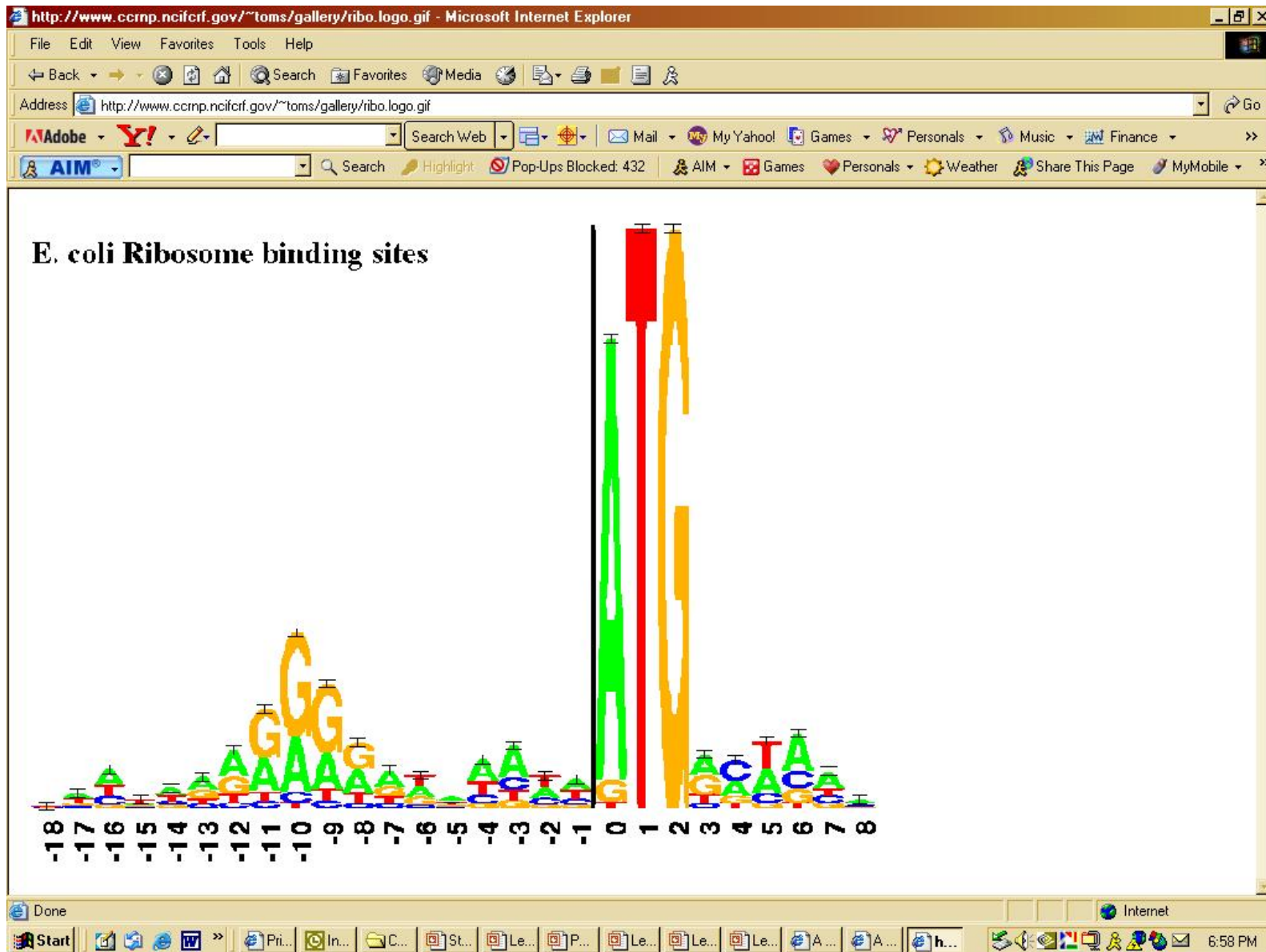
Testing of understanding

- What would maximal entropy of a column in multiple alignment of protein sequences?
- What is maximal /minimal information content of a column in multiple alignment of **protein** sequences?

Application: Finding functional residues

- Functionally important sites are expected to have higher than average information content (under assumption of common background this is equivalent to low entropy)

Sequence Logo



Question:

Does it always make sense to take uniform distribution as the as the random distribution while computing information content — why yes or why no?

Relative entropy

- Given are two probability distributions R and P.
Relative entropy of P to Q:

$$H(P \parallel Q) = \sum_{i=1..n} p_i \log (p_i / q_i) = \\ \sum_{i=1..n} p_i \log (p_i) - \sum_{i=1..n} p_i \log (q_i)$$

- Relative entropy is not a distance in mathematical sense (it is not symmetric).
- Relative entropy is closely related but not identical to information content:

$$I(P) = H(R) - H(P) = \sum_{i=1..n} p_i \log (p_i) - \sum_{i=1..n} q_i \log (q_i)$$

Example 1

Assume that the probability distribution of {A,T,C,G} at some column position in one sequence family is $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$ and in another family this distribution is: $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$ (uniformly distributed position). What is relative entropy?

$$\begin{aligned} H(P \parallel Q) &= \sum_{i=1..n} p_i \log(p_i / q_i) = \frac{1}{2} \log((1/2) / (1/4)) + \\ &\frac{1}{4} \log((1/4) / (1/4)) + \frac{1}{8} \log((1/8) / (1/4)) + \\ &\frac{1}{8} \log((1/8) / (1/4)) = \frac{1}{2} \log 2 + \frac{1}{4} \log 1 + \frac{1}{8} \\ &\log(1/2) + \frac{1}{8}(\log 1/2) = \frac{1}{2} + 0 + \frac{1}{8}(-1) + \frac{1}{8}(-1) = \\ &\frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \end{aligned}$$

Example 2

$$P = p_1, p_2, \dots, p_n; Q = q_1, q_2, \dots, q_n$$

Where

$$p_i = q_i$$

$$H(P||Q) = \sum_{i=1..n} p_i \log (p_i / q_i) = \sum_{i=1..n} p_i \log 1 = 0$$

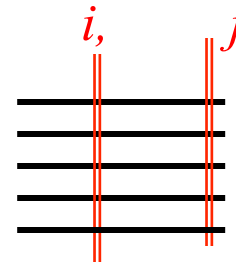
Mutual information

- X, Y random dependent variables. How much of information about X is contained in Y?
- Let **joint distribution** be $p(x,y)$ (i.e the probability distribution of joint occurrence of events X & Y) and **marginal distribution** $p(x)$, $p(y)$ (probability of observing X and Y separately)
- Mutual information context is the relative entropy between the joint distribution and the product distribution:

$$MI(X;Y) = H(p(x,y) \parallel p(x)p(y)) = \\ \sum_x \sum_y p(x,y) \log (p(x,y)/ p(x)p(y))$$

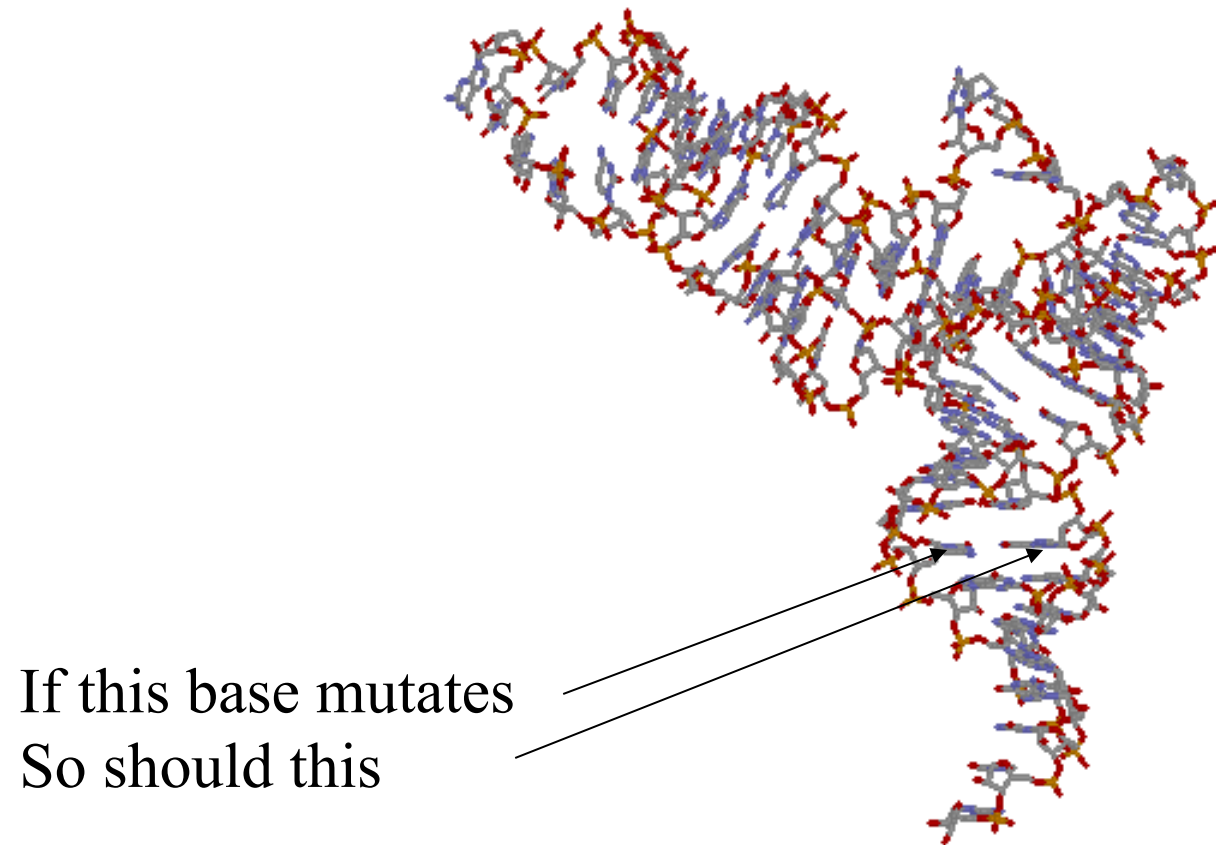
Example

- Assume that we are looking at two positions i and j in a set of aligned sequences



- In each column we observe some variations due to evolutionary changes
- Questions: Are changes in each column independent? Why do we care?
- If the changes are correlated then one this may suggest that the corresponding position are close in space

Correlated mutations in RNA



Bacillus subtilis RNase P RNA

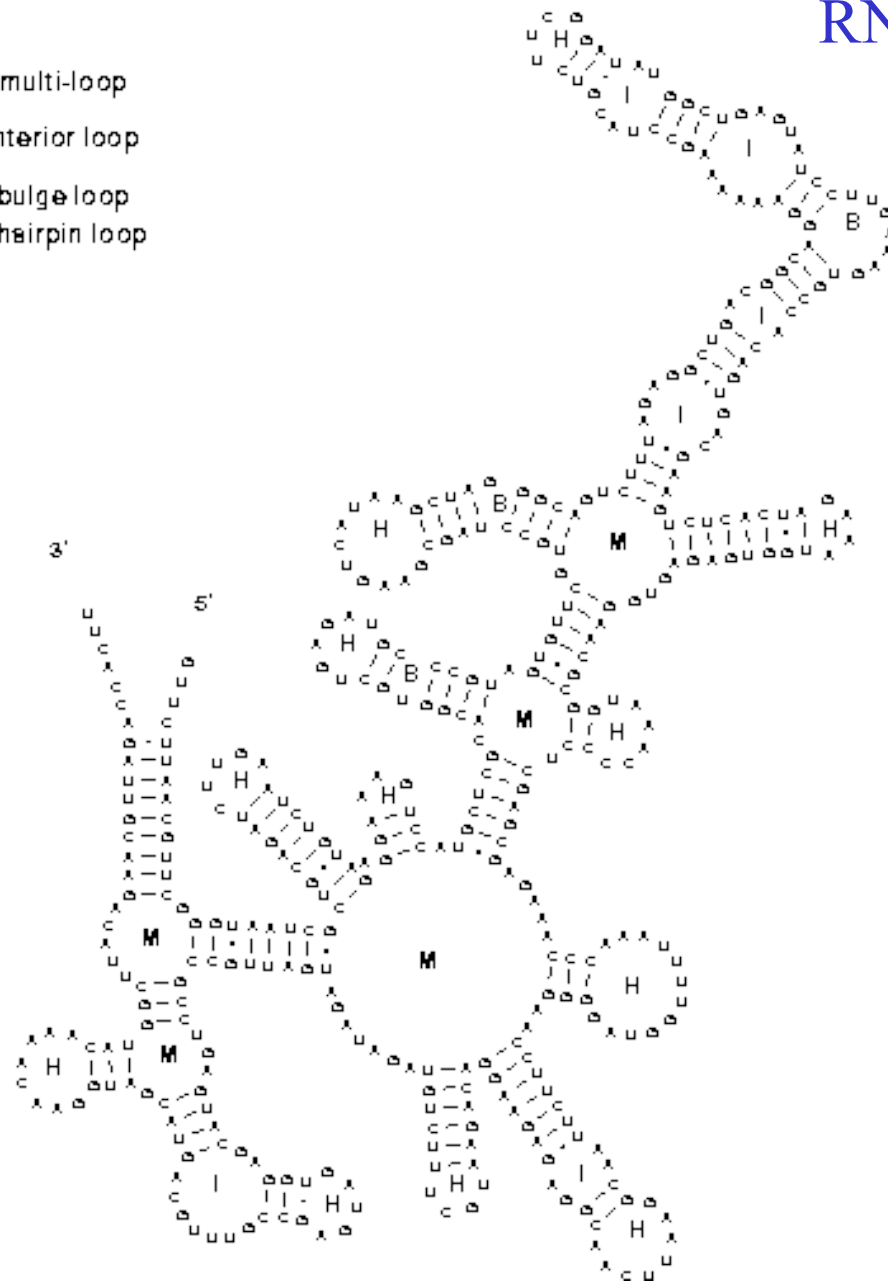
RNA secondary structure

M - multi-loop

I - interior loop

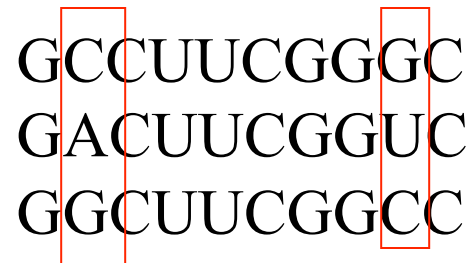
B - bulge loop

H - hairpin loop



Covariance method

- In a correct multiple alignment RNAs, conserved base pairs are often revealed by the presence of frequent correlated compensatory mutations,



GCCUUCGGGC
GACUUCGGUC
GGCUUCGGCC

The diagram shows three RNA sequences aligned. Two vertical red boxes highlight specific positions: the second position (C in the first sequence, A in the second, G in the third) and the eighth position (G in the first, G in the second, C in the third). These positions are boxed to illustrate compensatory mutations that maintain base pairing across the sequences.

Two boxed positions are **co-varying** to maintain Watson-Crick complementary. This covariation implies a base pair which may be then extended in both directions.

Example continued:

- In the information theoretic terms: how information in one column determine the information in the the other column

Examples

A
A
C
G

U
U
G
C

$$\begin{aligned}
 p(A) &= .5 \\
 p(C) &= .25 \\
 p(G) &= .25 \\
 p(U) &= .5 \\
 p(G) &= .25 \\
 p(C) &= .25
 \end{aligned}$$

$$\begin{aligned}
 p(A,U) &= .5 \\
 p(C,G) &= .25 \\
 p(G,C) &= .25
 \end{aligned}$$

Marginal probabilities

$$\begin{aligned}
 \sum_x \sum_y p(x,y) \log (p(x,y)/ p(x)p(y)) &= \\
 .5 \log_2 (.5/(.5*.5)) + 2*.25 \log_2 (.25/(.25*.25)) &= \\
 .5 * 1 + .5 * 2 &= 1.5
 \end{aligned}$$

A
A
A
A

U
U
U
U

$$M_{ij} = 1 \log 1 = 0$$

U
A
C
G

A
U
G
C

$$M_{ij} = 4*.25 \log 4 = 2$$

Applications of Mutual Information

- Predicting RNA structure
- Predicting interacting residues within protein
- Predicting functionally linked genes based on expression pattern